An Accelerated Inexact Dampened Augmented Lagrangian Method for Linearly-Constrained Nonconvex Composite Optimization Problems*

Weiwei Kong† and Renato D.C. Monteiro‡

August 12, 2022 (v1: October 23, 2021)

Abstract

This paper proposes and analyzes an accelerated inexact dampened augmented Lagrangian (AIDAL) method for solving linearly-constrained nonconvex composite optimization problems. Each iteration of the AIDAL method consists of: (i) inexactly solving a dampened proximal augmented Lagrangian (AL) subproblem by calling an accelerated composite gradient (ACG) subroutine; (ii) applying a dampened and under-relaxed Lagrange multiplier update; and (iii) using a novel test to check whether the penalty parameter of the AL function should be increased. Under several mild assumptions involving the dampening factor and the under-relaxation constant, it is shown that the AIDAL method generates an approximate stationary point of the constrained problem in $\mathcal{O}(\varepsilon^{-5/2}\log\varepsilon^{-1})$ iterations of the ACG subroutine, for a given tolerance $\varepsilon > 0$. Numerical experiments are also given to show the computational efficiency of the proposed method.

1 Introduction

This paper presents an accelerated inexact dampened augmented Lagrangian (AIDAL) method for finding approximate stationary points of the linearly constrained nonconvex composite optimization (NCO) problem

$$\min_{z}\left\{\phi(u) := f(z) + h(z) : Az = b\right\}, \tag{1}$$

where A is a linear operator, h is a proper closed convex and Lipschitz continuous function with compact domain, and f is a (possibly) nonconvex differentiable function on the domain of h with a Lipschitz continuous gradient. More specifically, the AIDAL method is based on the θ-dampened augmented Lagrangian (AL) function

$$\mathcal{L}_c^{\theta}(z;p) := \phi(z) + (1-\theta)\left\langle p, Az-b\right\rangle + \frac{c}{2}\|Az-b\|^2 \quad \forall c>0, \quad \forall \theta \in (0,1), \tag{2}$$

*The first author has been supported by (i) the US Department of Energy (DOE) and UT-Battelle, LLC, under contract DE-AC05-00OR22725, (ii) the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration, and (iii) the IDEaS-TRIAD Fellowship (NSF Grant CCF-1740776). The second author was partially supported by ONR Grant N00014-18-1-2077 and AFOSR Grant FA9550-22-1-0088.

†Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37830. wwkong92@gmail.com

‡School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0205. monteiro@isye.gatech.edu

and it performs the following updates to generate its k^{th} iterate: given (z_{k-1}, p_{k-1}) and (λ, c_k), compute

$$z_k \approx \underset{u}{\operatorname{argmin}} \left\{ \lambda \mathcal{L}_{c_k}^{\theta}(u; p_{k-1}) + \frac{1}{2}\|u - z_{k-1}\|^2 \right\}, \tag{3}$$

$$p_k = (1-\theta)p_{k-1} + \chi c_k (Az_k - b), \tag{4}$$

where χ is an under-relaxation parameter in $(0,1)$ and z_k is a suitably chosen approximate solution of the composite problem underlying (3). In addition, the AIDAL method introduces a novel approach for updating the penalty parameter c_k between iterations and uses an accelerated composite gradient (ACG) method applied to (3) obtain the aforementioned point z_k.

Under a suitable choice of λ and the following Slater-like assumption:

$$\exists \bar{z} \in \operatorname{int}(\operatorname{dom} h) \text{ such that } A\bar{z} = b, \tag{5}$$

where $\operatorname{int}(\operatorname{dom} h)$ denotes the interior of the domain of h, it is shown that, for any tolerance pair $(\rho, \eta) \in \mathbb{R}^2_{++}$, the AIDAL method obtains a pair $([\hat{z}, \hat{p}], \hat{v})$ satisfying

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + A^*\hat{p}, \quad \|\hat{v}\| \le \rho, \quad \|A\hat{z} - b\| \le \eta. \tag{6}$$

in $\mathcal{O}((\eta^{-5/2} + \eta^{-1/2}\rho^{-2}) \log \eta^{-1})$ ACG iterations. Moreover, this iteration complexity is obtained without requiring that the initial point z_0 (in the domain of h) be feasible with respect to the linear constraint, i.e., $Az_0 = b$. Another contribution from this analysis is that the sequence of Lagrange multipliers is shown to be bounded by a constant independent of ρ and η.

Related Works. To condense our discussion, we let $\varepsilon = \rho = \eta$ denote a common tolerance parameter and restrict our attention to works that establish iteration complexity bounds for obtaining approximate stationary points of (1). For an overview of papers that focus on asymptotic convergence of a proposed method, see the excellent discussion in [20, Section 2].

One popular class of methods for obtaining stationary points of (1) is the penalty method, which consists of solving a sequence of unconstrained subproblems containing an objective function that penalizes a violation of the constraints through a positively weighted penalty term. Papers [10, 15] present an $\mathcal{O}(\varepsilon^{-3})$ iteration complexity of a quadratic penalty method without any regularity assumptions on the linear constraint. In a follow-up work, paper [12] presents an $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ iteration complexity of a similar quadratic penalty method in which its parameters are chosen in an adaptive and numerically efficient manner. Paper [20] is the first to present a penalty-based method with an improved complexity of $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ under the assumption that the domain of h is compact and assumption (5) holds.

Another popular class of methods is the proximal AL (PAL) method, which primarily consists of the updates in (3) and (4). The analysis of AL/PAL-based methods for the case where ϕ is convex is already well-established (see, for example, [1, 2, 16, 17, 21, 22, 25, 26, 29]), so we make no more mention of it here. Instead, we review papers that present an iteration complexity of an AL/PAL-based method for the case where ϕ is nonconvex. Paper [6] presents an $\mathcal{O}(\varepsilon^{-4})$ iteration complexity[1] of an unaccelerated PAL method under the strong assumption that the initial point z_0 is feasible, i.e., $Az_0 = b$, as well as $\theta \in (0, 1]$ and $\chi = 1$. Paper [23] presents $\mathcal{O}(\varepsilon^{-3} \log \varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-5/2} \log \varepsilon^{-1})$ iteration complexities of an accelerated inexact PAL method for the general case and the case where (5) holds, respectively, and removes the requirement that the initial point

[1] This method generates prox subproblems of the form $\operatorname{argmin}_{x \in X}\{\lambda h(x) + c\|Ax - b\|^2/2 + \|x - x_0\|^2/2\}$ and the analysis of [6] makes the strong assumption that they can be solved exactly for any x_0, c, and λ.

be feasible. Papers [13, 14] present an $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$ iteration complexity for the special case of $(\chi, \theta) = (1, 0)$, which corresponds to a full multiplier update under the classical AL function. Finally, papers [28] and [18] respectively establish $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$ and $\mathcal{O}(\varepsilon^{-5/2}\log\varepsilon^{-1})$ iteration complexities for nonproximal AL-based methods that perform under-relaxed Lagrange multiplier updates only when the penalty parameter is updated.

Aside from penalty and AL/PAL-based methods, we mention few others that are of interest. Paper [3] presents an $\mathcal{O}(\varepsilon^{-3})$ iteration complexity of a primal-dual proximal point scheme for generating a point *near* an approximate stationary point under some strong conditions on the initial point. Papers [30,31] present an $\mathcal{O}(\varepsilon^{-2})$ iteration complexity of a primal-dual first-order algorithm for solving (1) when h is the indicator function of a box (in [31]), or more generally, a polyhedron (in [30]). Paper [7] presents an $\mathcal{O}(\varepsilon^{-6})$ iteration complexity of a penalty-ADMM method that solves an equivalent reformulation of (1), under the assumption that the initial point z_0 is feasible, the tolerance ε is sufficiently small, and A has full row rank. Paper [19] presents an inexact proximal point method applied to the function defined as $\phi(z)$ if z is feasible and $+\infty$ otherwise. It can be viewed as an extension to the nonconvex setting of the proximal point method (PPM) applied to (1) and it obtains an $\mathcal{O}(\varepsilon^{-5/2}\log\varepsilon^{-1})$ complexity bound.

Contributions. We now emphasize how the proposed AIDAL method improves on other state-of-the-art AL-based works. First, it improves upon the $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$ classic PAL method in [14] by an $\mathcal{O}(\varepsilon^{-1/2})$ factor through only a *small* perturbation of the classical multiplier update and the classical AL function. Second, AIDAL chooses its prox stepsize λ independent of the perturbation parameter θ. This is in contrast to the PAL method in [23] which has the undesirable property that its prox stepsize λ becomes arbitrarily small as θ approaches zero. Finally, it differs from the nonproximal AL-based method in [18] in two significant ways: (i) it performs the multiplier update (4) after every inexact prox update as opposed to only when the penalty parameter is updated; and (ii) it chooses a constant under-relaxation parameter χ for the update (4) as opposed to [18], which chooses an under-relaxation parameter that (linearly) tends to zero as the number of penalty parameter updates increases.

Organization of the Paper. Subsection 1.1 provides some basic definitions and notation. Section 2 contains two subsections. The first one describes the main problem of interest and the assumptions made on it, while the second one presents the AIDAL method and states its iteration complexity. Section 3 is divided into two subsections. The first one presents fundamental properties about two important residuals, while the second one gives the proof of a key proposition in Section 2. Section 4 presents numerical experiments that demonstrate the efficiency of the AIDAL method. Section 5 gives some concluding remarks. Finally, the end of the paper contains several important technical appendices.

1.1 Basic Notations and Definitions

This subsection presents notation and basic definitions used in this paper.

Let $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the set of nonnegative and positive real numbers, respectively, and let $\mathbb{R}^n$ denote the n-dimensional Hilbert space with inner product and associated norm denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$, respectively. The smallest positive singular value of a nonzero linear operator $Q : \mathbb{R}^n \to \mathbb{R}^l$ is denoted by σ_Q^+. For a given closed convex set $X \subset \mathbb{R}^n$, its boundary is denoted by ∂X and the distance of a point $x \in \mathbb{R}^n$ to X is denoted by $\mathrm{dist}_X(x)$. For any $t > 0$, we let $\log_1^+(t) := \max\{\log t, 1\}$ and denote $\mathcal{O}_1 = \mathcal{O}(\cdot + 1)$.

The domain of a function $h : \mathbb{R}^n \to (-\infty, \infty]$ is the set $\mathrm{dom}\, h := \{x \in \mathbb{R}^n : h(x) < +\infty\}$.

Moreover, h is said to be proper if $\operatorname{dom} h \neq \emptyset$. The set of all lower semi-continuous proper convex functions defined in $\mathbb{R}^n$ is denoted by $\overline{\text{Conv}}\ \mathbb{R}^n$. The subdifferential of a proper convex function $h : \mathbb{R}^n \to (-\infty, \infty]$ is defined by

$$\partial h(z) := \{u \in \mathbb{R}^n : h(z') \geq h(z) + \langle u, z' - z \rangle, \quad \forall z' \in \mathbb{R}^n\} \tag{7}$$

for every $z \in \mathbb{R}^n$. The normal cone of a closed convex set C at $z \in C$ is defined as

$$N_C(z) := \{\xi \in \mathbb{R}^n : \langle \xi, u - z \rangle \leq 0, \quad \forall u \in C\}.$$

If $\psi : \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable at $\bar{z} \in \mathbb{R}^n$, then its affine approximation at $\bar{z}$ is given by

$$\ell_\psi(z; \bar{z}) := \psi(\bar{z}) + \langle \nabla \psi(\bar{z}), z - \bar{z} \rangle \quad \forall z \in \mathbb{R}^n. \tag{8}$$

2 Augmented Lagrangian Method

This section contains two subsections. The first one precisely describes the problem of interest and the assumptions underlying it, while the second one presents the AIDAL method and its corresponding iteration complexity.

2.1 Problem of Interest

This subsection presents the main problem of interest and the assumptions underlying it.

Our problem of interest is precisely (1) where f, h, A, and b are assumed to satisfy the following assumptions:

(A1) $h \in \overline{\text{Conv}}\ \mathbb{R}^n$ is K_h-Lipschitz continuous and $\mathcal{H} := \operatorname{dom} h$ is compact with diameter $D_h := \sup_{u,z \in \mathcal{H}} \|u - z\| < \infty$.

(A2) f is differentiable function on $\mathcal{H}$, and there exists $(m, M) \in \mathbb{R}^2_{++}$ satisfying $m \leq M$, such that for every $u, z \in \mathcal{H}$, we have

$$-\frac{m}{2}\|u - z\|^2 \leq f(u) - f(z) - \langle \nabla f(z), u - z \rangle \leq \frac{M}{2}\|u - z\|^2, \tag{9}$$

$$\|\nabla f(u) - \nabla f(z)\| \leq M\|u - z\|; \tag{10}$$

(A3) there exists $\bar{z} \in \operatorname{int} \mathcal{H}$ such that $A\bar{z} = b$;

(A4) $A \neq 0$, $\mathcal{F} := \{z \in \mathcal{H} : Az = b\} \neq \emptyset$, and $\inf_{z \in \mathbb{R}^n} \phi(z) > -\infty$.

We now make four remarks about the above assumptions. First, it is well-known that (10) implies that $|f(u) - \ell_u(u; z)| \leq M\|u - z\|^2/2$ for every $u, z \in \mathcal{H}$ and hence that (9) holds with $m = M$. However, we show that better iteration complexities can be derived when a scalar $m < M$ satisfying (9) is available. Second, (9) implies that the function $f + m\|\cdot\|^2/2$ is convex on $\mathcal{H}$. Third, since $\mathcal{H}$ is compact by (A1), the image of any continuous $\mathbb{R}^k$-valued function, e.g., $u \mapsto \nabla f(u)$, is bounded. Finally, in Appendix C, we show that if $\hat{z}$ is a local minimum of (1), then there exists a multiplier $\hat{p}$ such that

$$0 \in \nabla f(\hat{z}) + \partial h(\hat{z}) + A^*\hat{p}, \quad A\hat{z} = b. \tag{11}$$

In view of the previous remark, we say that a pair $([\hat{z}, \hat{p}], \hat{v})$ is a (ρ, η)-stationary point of (1) if it satisfies condition (6), which is clearly a relaxation of (11) for any $(\rho, \eta) \in \mathbb{R}^2_{++}$.

2.2 AIDAL Method

This section presents the AIDAL method and its corresponding iteration complexity.

We first state the AIDAL method in Algorithm 2.1. Its main steps are: (i) invoking an ACG algorithm to apply the update in (3); (ii) computing a "refined" pair $(\hat{p}, \hat{v}) = (\hat{p}_k, \hat{v}_k)$ and point z satisfying the inclusion and (possibly) the inequality in (6); (iii) performing a novel test to determine the next penalty parameter c_{k+1}; and (iv) applying the update in (4).

Algorithm 2.1: Accelerated Inexact Dampened Augmented Lagrangian (AIDAL) Method

---

Input : $(m, M) \in \mathbb{R}^2_{++}$ as in (A2), $(\rho, \eta) \in \mathbb{R}^2_{++}$, $(z_0, p_0) \in \mathcal{H} \times A(\mathbb{R}^n)$, $c_1 \in \mathbb{R}_{++}$, $\sigma \in (0, 1/2]$, and $(\chi, \theta) \in (0,1)^2$ satisfying

$$2(1-\theta)(2-\theta)\chi \le \theta^2. \tag{12}$$

Output: a triple $(\hat{z}, \hat{p}, \hat{v}) \in \mathcal{H} \times A(\mathbb{R}^n) \times \mathbb{R}^n$ satisfying (6).

Function `AIDAL`$(\{m, M\}, \{\sigma, \chi, \theta\}, \{c_1, z_0, p_0\}, \{\rho, \eta\})$**:**

- $\lambda \leftarrow 1/(2m)$
- **for** $k \leftarrow 1, 2, \ldots$ **do**
  - `STEP 1` (inexact prox update): ▷ Implement (3)
  - $L_k \leftarrow \lambda(M + c_k\|A\|^2) + 1$
  - $\psi_s^k(\cdot) \leftarrow \lambda\left[\mathcal{L}^\theta_{c_k}(\cdot; p_{k-1}) - h(\cdot)\right] + \frac{1}{2}\|\cdot - z_{k-1}\|^2$ ▷ See (2) for the definition of $\mathcal{L}^\theta_c(\cdot;\cdot)$
  - $(z_k, v_k) \leftarrow$ `ACG`$(\{\psi_s^k, \lambda h\}, \{L_k, \frac{1}{2}\}, \sigma, z_{k-1})$ ▷ See Algorithm B.1
  - `STEP 2` (termination check):
  - $\hat{v}_k \leftarrow \frac{1}{\lambda}\left[v_k + z_{k-1} - z_k\right]$
  - $\hat{p}_k \leftarrow (1-\theta)p_{k-1} + c_k(Az_k - b)$
  - **if** $\|\hat{v}_k\| \le \rho$ **and** $\|Az_k - b\| \le \eta$ **then**
    - **return** $(z_k, \hat{p}_k, \hat{v}_k)$ ▷ Stop and output
  - `STEP 3` (penalty parameter update):
  - $c_{k+1} \leftarrow \begin{cases} 2c_k, & \text{if } \|\hat{v}_k\| \le \rho \text{ and } \|Az_k - b\| > \eta, \\ c_k, & \text{otherwise} \end{cases}$
  - `STEP 4` (multiplier update): ▷ Implement (4)
  - $p_k \leftarrow (1-\theta)p_{k-1} + \chi c_k(Az_k - b)$

Some remarks about Algorithm 2.1 are in order. First, its input z_0 can be any element in $\mathcal{H}$ and does not necessarily need to be a feasible point, i.e., one satisfying $Az_0 = b$. Second, its steps 1 and 4 are respectively the updates (3) and (4), while its step 3 consists of a test to determine whether the penalty parameter c_k should be increased. In particular, the update for (3) is obtained by applying the ACG algorithm in Algorithm B.1 to the (convex) proximal subproblem

$$\min_{u \in \mathbb{R}^n} \left\{ \lambda \mathcal{L}^\theta_{c_k}(\cdot; p_{k-1}) + \frac{1}{2}\|\cdot - z_{k-1}\|^2 \right\}$$

with an inexactness criterion (see (39)) that is a variant of the one considered by the authors in [9, 10, 13, 15]. Third, it performs two kinds of iterations: (i) the ones indexed by k; and (ii) the ones performed by the ACG algorithm every time it is called in its Step 1. To be concise, the former will be referred to as "outer" iterations and the latter as "inner" (or ACG) iterations. Finally, the computations in its step 2 are performed to ensure that the its output triple aligns with the notion of a (ρ,η)-approximate stationary point from the previous subsection (see Lemma 3.2 for more details).

We now present the key properties of the method. To be concise, we introduce the constants

$$\begin{gathered}\bar{d} := \mathrm{dist}_{\partial\mathcal{H}}(\bar{z}), \quad G_f := \sup_{u\in\mathcal{H}} \|\nabla f(u)\|, \quad \phi_* := \inf_{u\in\mathbb{R}^n} \phi(u), \quad \phi^* := \inf_{u\in\mathcal{F}} \phi(u), \\ \beta_\lambda := \left(\bar{d} + D_h\right)\left[K_h + G_f + \frac{(1+\sigma)D_h}{\lambda}\right],\end{gathered} \tag{13}$$

where $(D_h, K_h, \mathcal{H})$, $\bar{z}$, and $\mathcal{F}$ are as in (A1), (A3), and (A4), respectively. Moreover, we let

$$\mathcal{C}_\ell := \left\{k \in \mathbb{N} : c_k = c_1 2^{\ell-1}\right\} \tag{14}$$

denote the ℓ^{th} *cycle* of AIDAL method and, for simplicity, if the AIDAL method terminates at iteration k then the indices of the last cycle do not extend past k.

The first result presents a bound on the sequence of Lagrange multipliers $\{p_k\}_{k\geq 0}$ computed in step 3 of AIDAL. Its proof, which is given in Appendix A, is a modification of the one given in [14] for the case where $(\theta,\chi) = (0,1)$.

Proposition 2.1. *Let $\{p_i\}_{i\geq 1}$ be generated by the AIDAL method. Then,*

$$\|p_k\| \leq \frac{\max\{\|p_0\|, \beta_\lambda\}}{\min\{1, \bar{d}\sigma_A^+(1-\theta)\}} =: B_p \quad \forall k \geq 1.$$

The next result, whose proof is the topic of Subsection 3.2, describes several properties of AIDAL, including a bound on the number of inner iterations performed in each outer iteration, a uniform bound on the size of all cycles, and its successful termination with the required approximate stationary point of (1).

Proposition 2.2. *Let $(\lambda, c_1, \chi, \theta, \rho, \eta)$ be as in AIDAL, and define the nonnegative scalars*

$$\begin{gathered}B_\Psi := \phi^* - \phi_* + \frac{D_h^2}{\lambda} + \left(\frac{2-\theta+2[2-\theta][1-\theta]}{2\chi^2 c_1}\right) B_p^2, \\ \bar{c}_\eta := \frac{2B_p}{\chi\eta}, \quad \mathcal{T}_\rho := \left\lceil 1 + \frac{32 B_\Psi}{\lambda\rho^2}\right\rceil.\end{gathered} \tag{15}$$

where B_p, D_h, and (ϕ_*, ϕ^*) are as in Proposition 2.1, assumption (A1), and (13), respectively. Then, the following statements hold about AIDAL:

(a) its k^{th} outer iteration performs a number of inner (or ACG) iterations bounded above by

$$\left\lceil 1 + 12\sqrt{L_k}\log_1^+ \frac{7L_k}{\sigma}\right\rceil;$$

(b) for every $\ell \geq 1$, it holds that $|\mathcal{C}_\ell| \leq \mathcal{T}_\rho$, and the residual $\hat{v}_k$ for the last index k of $\mathcal{C}_\ell$ satisfies $\|\hat{v}_k\| \leq \rho$;

(c) the last cycle $\bar{\ell}$ outputs a (ρ,η)-stationary point of (1) and satisfies $c_k \leq \max\{c_1, 2\bar{c}_\eta\}$ for every $k \in \mathcal{C}_{\bar{\ell}}$; as a consequence, $\bar{\ell} \leq \max\{1, \log_2(2\bar{c}_\eta/c_1)\}$.

We give some remarks about the above results. First, Proposition 2.1 states that the sequence of Lagrange multipliers $\{p_k\}_{k\geq 1}$ generated by the AIDAL method is bounded by a constant that is independent of the tolerances ρ and η. Second, Proposition 2.2(c) states that the number of times that the penalty constant c_k is doubled during an invocation of the AIDAL method is finite. Finally, Proposition 2.2(a) shows that the number of the inner (or ACG) iterations at each outer iteration of AIDAL is independent of the tolerances ρ and η.

Using Proposition 2.2, the next result establishes an $\mathcal{O}(\eta^{-1/2}\rho^{-2}\log\eta^{-1})$ total inner (or ACG) iteration complexity for the AIDAL method.

Theorem 2.3. *AIDAL stops with a (ρ,η)-stationary point of (1) in a number of inner (or ACG) iterations bounded above by*

$$\mathcal{O}_1\left(\mathcal{T}_\rho\sqrt{\bar{c}_\eta L_1}\log_1^+\frac{\bar{c}_\eta L_1}{\sigma}\right), \tag{16}$$

where $(\bar{c}_\eta,\mathcal{T}_\rho)$ are as (15), L_1 is as in step 1 of AIDAL, and σ is the inexactness parameter given to AIDAL.

Proof. For ease of notation, let $(\bar{c},\mathcal{T})=(\bar{c}_\eta,\mathcal{T}_\rho)$. In view of Proposition 2.2(a) and (c), the total number of inner (or ACG) iterations performed by the method is on the order of

$$\mathcal{O}_1\left(\sum_{\ell=1}^{\lceil\log_2\bar{c}\rceil}\sum_{j\in\mathcal{C}_\ell}\sqrt{L_j}\log_1^+\frac{L_j}{\sigma}\right). \tag{17}$$

To simplify this sum, we first note that if $c_j\in\mathcal{C}_\ell$, then the relations $\lambda=1/m$ (from AIDAL) and $m\leq M$ (from assumption (A2)) imply that

$$L_j=\lambda\left(M+2^{\ell-1}c_1\|A\|^2+\lambda^{-1}\right)\leq\lambda\left(2M+2^{\ell-1}c_1\|A\|^2\right)\leq 2^\ell L_1. \tag{18}$$

Combining (18) with Proposition 2.2(b), it holds that

$$\begin{aligned}\sum_{\ell=1}^{\lceil\log_2\bar{c}\rceil}\sum_{j\in\mathcal{C}_\ell}\sqrt{L_j}&\leq\mathcal{T}\sqrt{L_1}\sum_{\ell=1}^{\lceil\log_2\bar{c}\rceil}2^{\ell/2}=\mathcal{T}\sqrt{L_1}\cdot\sqrt{2}\left(1+\sqrt{2}\right)\left(2^{\lceil\log_2\bar{c}\rceil/2}-1\right)\\&\leq 4\mathcal{T}\sqrt{L_1}\left(2^{\log_2\sqrt{\bar{c}}}\cdot 2^{1/2}\right)=\mathcal{O}_1\left(\mathcal{T}\sqrt{\bar{c}L_1}\right).\end{aligned} \tag{19}$$

Moreover, denoting $\bar{\ell}=\lceil\log_2\bar{c}\rceil$, it follows from (18) that

$$\max_{1\leq\ell\leq\lceil\log_2\bar{c}\rceil}\max_{j\in\mathcal{C}_\ell}\left\{\log_1^+L_j\right\}=\log_1^+\left[\lambda\left(M+c_{\bar{\ell}}A\|^2\right)+1\right]=\mathcal{O}_1\left(\log_1^+[\bar{c}L_1]\right). \tag{20}$$

The complexity bound in (16) now follows from (19), (20), and (17). The fact that AIDAL stops with a (ρ,η)-stationary point of (1) follows from Proposition 2.2(c). $\square$

Notice that if $B_\Psi=\Theta(1)$ then, in terms of only the tolerance pair (ρ,η) and the stepsize λ, Theorem 2.3 states that the AIDAL outputs a (ρ,η)-approximate stationary point of (1) in

$$\mathcal{O}_1\left(\frac{1}{\sqrt{\eta}}\left[\sqrt{\lambda}+\frac{1}{\rho^2\sqrt{\lambda}}\right]\log_1^+\frac{\lambda}{\eta}\right) \tag{21}$$

inner (or ACG) iterations. Also, the number of resolvent (or proximal) evaluations of h in AIDAL is on the same order of magnitude as in (16) from the fact that the ACG algorithm in Appendix B performs only one resolvent evaluation every ACG iteration.

3 Convergence Analysis of the AIDAL Method

This section establishes the key properties of the AIDAL method and contains two subsections. The first one presents fundamental properties about two key residuals, while the second gives the proof of Proposition 2.2.

To avoid repetition, we let

$$\{(z_i, p_i, v_i, \hat{p}_i, \hat{v}_i, \psi_s^i, c_i, L_i)\}_{i\geq 1},$$

denote the sequence of iterates generated by the AIDAL method. Moreover, for every $i \geq 1$ and any $(\chi, \theta) \in \mathbb{R}^2_{++}$, we make use of the following useful constants

$$\begin{gathered} a_\theta = \theta(1-\theta), \quad b_\theta := (2-\theta)(1-\theta), \quad \alpha_{\chi,\theta} := \frac{(1-2\chi b_\theta)-(1-\theta)^2}{2\chi}, \\ f_i := Az_i - b, \quad \Delta p_i = p_i - p_{i-1}, \quad \Delta z_i = z_i - z_{i-1}, \end{gathered} \tag{22}$$

and the iterate-dependent potential

$$\Psi_i^\theta := \mathcal{L}_{c_i}^\theta(z_i; p_i) - \frac{a_\theta}{2\chi c_i}\|p_i\|^2 + \frac{\alpha_{\chi,\theta}}{4\chi c_i}\|\Delta p_i\|^2. \tag{23}$$

3.1 Fundamental Properties about Key Residuals

This subsection focuses on establishing the following result, which presents some fundamental properties about the residuals $\hat{v}_i$ and f_i.

Proposition 3.1. *The following statements hold:*

(a) for every $i \geq 1$, we have

$$\hat{v}_i \in \nabla f(z_i) + \partial h(z_i) + A^*\hat{p}_i, \quad \|f_i\| \leq \frac{2B_p}{\chi c_i};$$

(b) for every $\ell \geq 1$ and $j, k \in \mathcal{C}_\ell$ such that $k \geq j+1$, we have $\lambda \sum_{i=j+1}^k \|\hat{v}_i\|^2 \leq 32[\Psi_j^\theta - \Psi_k^\theta]$.

We now make a few comments about the above statements. First, in view of the inclusion of Proposition 3.1(a), it follows that $(z_i, \hat{p}_i, \hat{v}_i)$ is a (ρ, η)-stationary point of (1) if and only if $\|\hat{v}_i\| \leq \rho$ and $\|f_i\| \leq \eta$. Second, the inequality of Proposition 3.1(b) implies that the feasibility residual $\|f_i\|$ can be made small by making the penalty parameter sufficiently large. Third, it will be shown in the next subsection that the potential Ψ_i^θ is bounded both from above and from below, which, together with Proposition 3.1(c), then implies that, within a cycle, a residual $\hat{v}_i$ such that $\|\hat{v}_i\| \leq \rho$ can be found in $O(\rho^{-2})$ iterations of the cycle.

We now present some basic properties about the iterates generated by the method.

Lemma 3.2. *For every $i \geq 1$, the following statements hold:*

(a) $f_i = [p_i - (1-\theta)p_{i-1}]/(\chi c_i)$;

(b) if $i \geq 2$, then $\chi(c_i f_i - c_{i-1} f_{i-1}) = \Delta p_i - (1-\theta)\Delta p_{i-1}$;

(c) it holds that

$$v_i \in \partial\left(\lambda \mathcal{L}_{c_i}^\theta(\cdot; p_{i-1}) + \frac{1}{2}\|\cdot - z_{i-1}\|^2\right)(z_i), \quad \|v_i\| \leq \sigma\|\Delta z_i\|, \quad \|\hat{v}_i\| \leq \frac{1+\sigma}{\lambda}\|\Delta z_i\|.$$

Proof. (a) This follows from the definition of f_i in (22), and step 4 of the AIDAL method.

(b) This follows from part (a) and the definition of Δp_i in (22).

(c) The inclusion and first bound follow from the ACG call in step 1 of the AIDAL method, Proposition B.1 with $(\psi_s, \psi_n) = (\psi_s^i, \lambda h)$ and $(z, v, x_0) = (z_i, v_i, z_{i-1})$, and the fact that $\lambda \mathcal{L}_{c_i}^\theta(\cdot; p_{i-1}) + \|\cdot - z_{i-1}\|^2/2$ is convex (see the choice of λ and assumption (A2)). For the second bound, we use the first bound, the triangle inequality, and the definition of $\hat{v}_i$ to obtain

$$\|\hat{v}_i\| = \frac{1}{\lambda}\|v_i + z_{i-1} - z_i\| \le \frac{1}{\lambda}\|v_i\| + \frac{1}{\lambda}\|\Delta z_i\| \le \frac{1+\sigma}{\lambda}\|\Delta z_i\|. \qquad \square$$

The next result gives a more refined bound on $\|\hat{v}_i\|$.

Lemma 3.3. *For every $i \ge 1$, it holds that*

$$\begin{aligned}\frac{\lambda}{32}\|\hat{v}_i\|^2 \le &\left[\mathcal{L}_{c_i}^\theta(z_{i-1}; p_{i-1}) - \mathcal{L}_{c_i}^\theta(z_i; p_i) + \frac{a_\theta}{2\chi c_i}\left(\|p_i\|^2 - \|p_{i-1}\|^2\right)\right] \\ &+ \frac{b_\theta}{2\chi c_i}\|\Delta p_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2,\end{aligned} \tag{24}$$

where a_θ and b_θ are as in (22).

Proof. Let $i \ge 1$ be fixed. We first derive a relationship for $\mathcal{L}_{c_i}^\theta(z_i, p_i) - \mathcal{L}_{c_i}^\theta(z_i, p_{i-1})$. Using the definition of $\mathcal{L}_c^\theta$ in (2), the definitions of Δp_i and f_i in (22), and Lemma 3.2(a), we have that

$$\begin{aligned}\mathcal{L}_{c_i}^\theta(z_i, p_i) - \mathcal{L}_{c_i}^\theta(z_i, p_{i-1}) &= (1-\theta)\langle \Delta p_i, f_i\rangle = \left(\frac{1-\theta}{\chi c_i}\right)\|\Delta p_i\|^2 + \frac{(1-\theta)\theta}{\chi c_i}\langle \Delta p_i, p_{i-1}\rangle \\ &= \left(\frac{1-\theta}{\chi c_i}\right)\|\Delta p_i\|^2 + \frac{(1-\theta)\theta}{\chi c_i}\left(\langle p_i, p_{i-1}\rangle - \|p_{i-1}\|^2\right) \\ &= \left(\frac{1-\theta}{\chi c_i}\right)\|\Delta p_i\|^2 + \frac{(1-\theta)\theta}{\chi c_i}\left(-\frac{1}{2}\|\Delta p_i\|^2 + \frac{1}{2}\|p_i\|^2 - \frac{1}{2}\|p_{i-1}\|^2\right) \\ &= \frac{b_\theta}{2\chi c_i}\|\Delta p_i\|^2 + \frac{a_\theta}{2\chi c_i}\left(\|p_i\|^2 - \|p_{i-1}\|^2\right).\end{aligned} \tag{25}$$

We next derive a bound for $\mathcal{L}_{c_i}^\theta(z_i, p_i) - \mathcal{L}_{c_i}^\theta(z_i, p_{i-1})$. Define the quantities

$$Q_{c_i} := (1-\lambda m)I + c_i\lambda A^*A, \quad \|x\|_{c_i} := \langle x, Q_{c_i}x\rangle \quad \forall x \in \mathbb{R}^n.$$

Moreover, observe that assumption (A2) and the definition of $\mathcal{L}_c^\theta$ imply that $\psi_s^i + \lambda h$ is (1/2)-strongly convex with respect to $\|\cdot\|_{c_i}$. As a consequence, it follows from the inclusion in Lemma 3.2(c) that

$$v_i \in \partial\left(\lambda\mathcal{L}_{c_i}^\theta(\cdot; p_{i-1}) + \frac{1}{2}\|\cdot - z_{i-1}\|^2 - \frac{1}{4}\|\cdot - z_i\|_{c_i}^2\right)(z_i).$$

which, in particular, implies that

$$\lambda\mathcal{L}_{c_i}^\theta(z_{i-1}; p_{i-1}) - \frac{1}{4}\|\Delta z_i\|_{c_i}^2 \ge \lambda\mathcal{L}_{c_i}^\theta(z_i; p_{i-1}) + \frac{1}{2}\|\Delta z_i\|^2 - \langle v_i, \Delta z_i\rangle. \tag{26}$$

Combining (26), the inequality in Lemma 3.2(c), the fact that $\sigma \in (0, 1/2]$ and $\lambda = 1/(2m)$, and the Cauchy-Schwarz inequality, we conclude that

$$\begin{aligned}
&\mathcal{L}^{\theta}_{c_i}(z_i,p_{i-1}) - \mathcal{L}^{\theta}_{c_i}(z_{i-1},p_{i-1}) \le -\frac{1}{4\lambda}\|\Delta z_i\|^2_{c_i} - \frac{1}{2\lambda}\|\Delta z_i\|^2 + \frac{1}{\lambda}\langle v_i, \Delta z_i\rangle \\
&= -\left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2 - \frac{1}{2\lambda}\left(\|\Delta z_i\|^2 - 2\langle v_i,\Delta z_i\rangle\right) \\
&\le -\left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2 - \frac{1}{2\lambda}\left(\|\Delta z_i\|^2 - 2\|v_i\|\cdot\|\Delta z_i\|\right) \\
&\le -\left(\frac{1-\lambda m}{4\lambda}\right)\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2 - \left(\frac{1-2\sigma}{2\lambda}\right)\|\Delta z_i\|^2 \\
&= -\left(\frac{3-\lambda m-4\sigma}{4\lambda}\right)\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2 \le \frac{1}{8\lambda}\|\Delta z_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2
\end{aligned} \tag{27}$$

The conclusion now follows by summing (25) and (27), isolating the $\|\Delta z_i\|^2$ term to one side, and using the inequality on $\|\hat{v}_i\|$ in Lemma 3.2 with the fact that $(1+\sigma)^2 \le 4$. □

Note that within a cycle, where the penalty parameters remain constant, the term within the square bracket of the right-hand side of (24) is telescopic. Interestingly, the next result shows that the other term on the right-hand side of (24) can be telescopically bounded within a fixed cycle. It is worth mentioning that the relationship between χ and θ in (12) plays an important role in proving this fact.

Lemma 3.4. *For every $i \ge 2$ such that $c_i = c_{i-1}$, it holds that*

$$\frac{b_\theta}{2\chi c_i}\|\Delta p_i\|^2 - \frac{c_i}{4}\|A\Delta z_i\|^2 \le \frac{\alpha_{\chi,\theta}}{4\chi c_i}\left(\|\Delta p_{i-1}\|^2 - \|\Delta p_i\|^2\right), \tag{28}$$

where b_θ and $\alpha_{\chi,\theta}$ are as in (22).

Proof. Let $i \ge 2$ be an index where $c_i = c_{i-1}$ and observe that (12) implies $2\chi b_\theta \le \theta^2$. Moreover, define

$$\widehat{\Delta} p_i := \Delta p_i - (1-\theta)\Delta p_{i-1}$$

and observe that Lemma A.1 with $(\tau, a, b) = (2\chi b_\theta, \Delta p_i, \Delta p_{i-1})$ implies that

$$\frac{1}{\chi}\|\widehat{\Delta} p_i\|^2 \ge 2b_\theta\|\Delta p_i\|^2 + \alpha_{\chi,\theta}\left(\|\Delta p_i\|^2 - \|\Delta p_{i-1}\|^2\right). \tag{29}$$

Using Lemma 3.2(b), the fact that $c_i = c_{i-1}$, and (29), we then have

$$\begin{aligned}
\frac{c_i}{4}\|A\Delta z_i\|^2 &= \frac{\|\chi c_i A\Delta z_i\|^2}{4\chi^2 c_i} = \frac{\|\chi(c_i f_i - c_{i-1} f_{i-1})\|^2}{4\chi^2 c_i} = \frac{1}{4\chi c_i}\left[\frac{1}{\chi}\|\widehat{\Delta} p_i\|^2\right] \\
&\ge \frac{1}{4\chi c_i}\left[2b_\theta\|\Delta p_i\|^2 + \alpha_{\chi,\theta}\left(\|\Delta p_i\|^2 - \|\Delta p_{i-1}\|^2\right)\right],
\end{aligned}$$

from which (28) immediately follows. □

Combining (24) and (28), it is easy to see that the sum of the residuals $\{\|\hat{v}_i\|^2\}_{i\ge 1}$ residuals is bounded above by a telescopic sum when the indices are in a cycle. Let us now use this fact with the bound on $\|f_i\|$ in Lemma 3.2 to prove Proposition 3.1.

Proof of Proposition 3.1. (a) It follows from the convexity of $\lambda\mathcal{L}^{\theta}_{c_i}(\cdot;p_{i-1})+\frac{1}{2}\|\cdot-z_{i-1}\|^2$ (see the choice of λ and assumption (A2)), Lemma 3.2(c), and the definition of $\hat{v}_i$ that

$$\begin{aligned}\hat{v}_i=\frac{v_i+z_{i-1}-z_i}{\lambda}&\in\frac{1}{\lambda}\partial\left(\lambda\mathcal{L}^{\theta}_{c_i}(\cdot;p_{i-1})+\frac{1}{2}\|\cdot-z_{i-1}\|^2\right)(z_i)+\frac{z_{i-1}-z_i}{\lambda}\\&=\nabla_z\mathcal{L}^{\theta}_{c_i}(z_i;p_{i-1})=\nabla f(z_i)+\partial h(z_i)+(1-\theta)A^*p_{i-1}+cA^*(Az_i-b)\\&=\nabla f(z_i)+\partial h(z_i)+A^*\hat{p}_i.\end{aligned}$$

On the other hand, using Lemma 3.2(a), the fact that $1-\theta\in[0,1]$, Proposition 2.1, and the triangle inequality, we have that for every $i\geq 1$,

$$\|f_i\|=\frac{\|p_i-(1-\theta)p_{i-1}\|}{\chi c_i}\leq\frac{\|p_i\|+(1-\theta)\|p_{i-1}\|}{\chi c_i}\leq\frac{2B_p}{\chi c_i}.$$

(b) Notice that for every $i\geq 2$ such that $c_i=c_{i-1}$, it follows from Lemmas 3.3 and 3.4 that

$$\begin{aligned}\frac{\lambda}{32}\|\hat{v}_i\|^2&\leq\mathcal{L}^{\theta}_{c_{i-1}}(z_{i-1};p_{i-1})-\mathcal{L}^{\theta}_{c_i}(z_i;p_i)+\frac{a_\theta}{2\chi c_i}\left(\|p_i\|^2-\|p_{i-1}\|^2\right)+\frac{\alpha_{\chi,\theta}}{4\chi c_i}\left(\|\Delta p_{i-1}\|^2-\|\Delta p_i\|^2\right)\\&=\Psi^{\theta}_{i-1}-\Psi^{\theta}_i.\end{aligned}$$

where the above identity is due to the definition of Ψ^{θ}_i. Noting that the assumption that $j,k\in\mathcal{C}_\ell$ implies that $c_i=c_{i-1}$ for every $i\in\{j+1,\ldots,k\}$, we conclude that (b) follows by summing the above inequality from $i=j+1$ to k. $\square$

3.2 Proof of Proposition 2.2

Proposition 3.1 plays an important role in the proof of Proposition 2.2. The first part of this subsection further refines Proposition 3.1(b) to show that the right-hand side of its inequality is $\mathcal{O}(1)$, a conclusion that follows from the fact that the potential Ψ^{θ}_i is bounded both from above and from below as will be shown below.

Lemma 3.5. *For every $i\geq 1$, it holds that*

$$\phi_*-\left(\frac{1-\theta}{2\chi c_1}\right)B_p^2\leq\Psi^{\theta}_i\leq\phi^*+\frac{D_h^2}{\lambda}+\left(\frac{1+2b_\theta}{2\chi^2 c_1}\right)B_p^2,\tag{30}$$

where (ϕ_*,ϕ^*), b_θ, and D_h are as in (13), (22), *and assumption (A1), respectively.*

Proof. Let $i\geq 1$. Using Proposition 2.1, the definitions of $\mathcal{L}^{\theta}_c(\cdot,\cdot)$, Ψ^{θ}_j, ϕ_*, and B_p, and the fact that $\chi\in(0,1)$, we have

$$\begin{aligned}\Psi^{\theta}_i&\geq\mathcal{L}^{\theta}_{c_i}(z_i;p_i)-\frac{a_\theta}{2\chi c_i}\|p_i\|^2=\phi(z_i)+(1-\theta)\langle p_i,Az_i-b\rangle+\frac{c_i}{2}\|Az_i-b\|^2-\frac{a_\theta}{2\chi c_i}\|p_i\|^2\\&\geq\phi_*+\frac{1}{2}\left\|\left(\frac{1-\theta}{\sqrt{c_i}}\right)p_i+\sqrt{c_i}(Az_i-b)\right\|^2-\frac{(1-\theta)^2}{2c_i}\|p_i\|^2-\frac{a_\theta}{2\chi c_i}\|p_i\|^2\\&\geq\phi_*-\left[\frac{(1-\theta)^2+a_\theta}{2\chi c_i}\right]B_p^2\geq\phi_*-\left(\frac{1-\theta}{2\chi c_1}\right)B_p^2,\end{aligned}$$

which is the desired lower bound in (30). For the upper bound, let an arbitrary $u \in \mathcal{F}$ be given. Using the fact that $Au = b$ and $u \in \mathcal{H}$, the definitions of $\mathcal{L}_c^\theta(\cdot,\cdot)$ and D_h , Lemma 3.2(c), and the Cauchy-Schwarz inequality, we conclude that

$$\begin{aligned}\lambda\mathcal{L}_{c_i}^\theta(z_i;p_{i-1}) &\overset{\text{Lemma 3.2}(c)}{\leq} \lambda\mathcal{L}_{c_i}^\theta(u;p_{i-1}) + \frac{1}{2}\|u-z_{i-1}\|^2 - \frac{1}{2}\|\Delta z_i\|^2 - \langle v_i, u - z_i\rangle \\ &\overset{u\in\mathcal{F}}{\leq} \lambda\phi(u) + \frac{1}{2}D_h^2 + \|v_j\|D_h \overset{\text{Lemma 3.2}(c)}{\leq} \lambda\phi(u) + \left(\frac{1}{2}+\sigma\right)D_h^2.\end{aligned}$$

Taking the infimum of the above bound over $u \in \mathcal{F}$ and using the fact that $\sigma \in (0, 1/2]$, we thus have $\mathcal{L}_{c_i}^\theta(z_i;p_{i-1}) \leq \phi^* + D_h^2/\lambda$. This inequality, (25), the fact that $\chi \in (0,1)$, Proposition 2.1, and the relation $(a+b)^2 \leq 2a^2 + 2b^2$ for every $a, b \in \mathbb{R}$, then imply that

$$\begin{aligned}\Psi_i^\theta &= \mathcal{L}_{c_i}^\theta(z_i;p_i) - \frac{a_\theta}{2\chi c_i}\|p_i\|^2 + \frac{\alpha_{\chi,\theta}}{4\chi c_i}\|\Delta p_i\|^2 \leq \mathcal{L}_{c_i}^\theta(z_i;p_{i-1}) + \left(\frac{2b_\theta+\alpha_{\chi,\theta}}{4\chi c_i}\right)\|\Delta p_i\|^2 \\ &\leq \phi^* + \frac{D_h^2}{\lambda} + \left(\frac{2b_\theta+\alpha_{\chi,\theta}}{2\chi c_i}\right)(\|p_i\|^2 + \|p_{i-1}\|^2) \leq \phi^* + \frac{D_h^2}{\lambda} + \left(\frac{1+2b_\theta}{2\chi^2 c_1}\right)B_p^2,\end{aligned}$$

which is the desired upper bound in (30). $\square$

Let us now combine the above bounds on Ψ_i^θ to obtain a more useful bound on $\hat{v}_i$.

Lemma 3.6. *For every $\ell \geq 1$ and $j, k \in \mathcal{C}_\ell$ such that $j < k$, there exists $i \in \{j+1, ..., k\}$ such that*

$$\lambda\|\hat{v}_i\|^2 \leq \frac{32B_\Psi}{k-j}, \tag{31}$$

where B_Ψ is as in (15).

Proof. Denote $\tilde{c}_\ell$ the constant value of c_j for every $j \in \mathcal{C}_\ell$ and let Ψ_i^θ be as in (23). Using the first bound of (30) with $i = k$ and the second bound of (30) with $i = j$, we first have that

$$\Psi_j^\theta - \Psi_k^\theta \leq \phi^* - \phi_* + \frac{D_h^2}{\lambda} + \left(\frac{2-\theta+2b_\theta}{2\chi^2 c_1}\right)B_p^2 = B_\Psi,$$

where B_Ψ is as in (15). Using the above bound and Proposition 3.1, it follows that

$$\lambda(k-j)\min_{j+1\leq i\leq k}\|\hat{v}_i\|^2 \leq \lambda\sum_{i=j+1}^{k}\|\hat{v}_i\|^2 \leq 32\left(\Psi_j^\theta - \Psi_k^\theta\right) \leq 32B_\Psi,$$

which implies the existence of some $i \in \{j+1, ..., k\}$ satisfying the bound on $\|\hat{v}_i\|$ in (31). $\square$

We are now ready to give the proof of Proposition 2.2.

Proof of Proposition 2.2. (a) Using the fact that $L_k \geq 1$, we first observe that for $\mu = 1/2$ we have

$$\frac{4L_k^2}{\mu}\left(\frac{1}{\mu} + \frac{36L_k}{\sigma^2}\right) = 8L_k^2\left(2 + \frac{36L_k}{\sigma^2}\right) \leq 8L_k^2\left(\frac{L_k}{\sigma^2} + \frac{36L_k}{\sigma^2}\right) \leq \frac{296L_k^3}{\sigma^2} \leq \left[\frac{7L_k}{\sigma}\right]^3.$$

The conclusion now follows from step 1 of the AIDAL method, assumption (A2), Proposition B.1(b) with $(L,\mu) = (L_k, 1/2)$, and the above bound.

(b) The fact that the last index k of a cycle $\mathcal{C}_\ell$ satisfies $\|\hat{v}_k\| \leq \rho$ follows immediately from steps 2–3 of AIDAL. Now, let $\ell \geq 1$ be fixed and define $j := \inf\{i : i \in \mathcal{C}_\ell\}$ and $k := j + \mathcal{T}_\rho - 1$. If $k \notin \mathcal{C}_\ell$

then $|\mathcal{C}_\ell| \le k - j + 1 = \mathcal{T}_\rho$. On the other hand, if $k \in \mathcal{C}_\ell$ then Lemma 3.6 and the definition of $\mathcal{T}_\rho$ in (15) imply that there exists $i \in \{j+1, ..., k\}$ such that

$$\|\hat{v}_i\|^2 \le \frac{32 B_\Psi}{\lambda(\mathcal{T}_\rho - 1)} \le \rho^2.$$

Since every cycle stops when $\|\hat{v}_i\| \le \rho$, we conclude that $i = k = \sup\{i : i \in \mathcal{C}_\ell\}$ and, hence, $|\mathcal{C}_\ell| = k - j + 1 = \mathcal{T}_\rho$.

(c) Let $\bar{c} = \bar{c}_\eta$. We first establish the bound on c_k. If AIDAL stops in the first cycle, then the bound on c_k follows immediately. Assume now that there is more than one cycle and suppose, for the sake of contradiction, that there exists a cycle $\ell \ge 2$ such that $c_k > 2\bar{c}$ for every $k \in \mathcal{C}_\ell$, and let k' denote the last index in $\mathcal{C}_{\ell-1}$. In view steps 3 of AIDAL, we then have $c_{k'} > \bar{c}$. Using the previous bound, the definition of $\bar{c} = \bar{c}_\eta$ in (15), Proposition 3.1(a), we also have

$$\|Ax_{k'} - b\| = \|f_{k'}\| \le \frac{2B_p}{\chi c_{k'}} \le \frac{2B_p}{\chi \bar{c}} \le \eta.$$

However, since $\|\hat{v}_{k'}\| \le \rho$ from part (b), this is impossible because termination would have occurred at the end of cycle $\ell - 1$. Hence, $c_k \le \max\{c_1, 2\bar{c}\}$. Since $c_k = 2^{\bar{\ell}-1} c_1$ for every $k \in \mathcal{C}_{\bar{\ell}}$, the bound on $\bar{\ell}$ is immediate. Moreover, it follows from parts (a)–(b) and the fact that $\bar{\ell}$ is finite that AIDAL always stops in step 2. Hence, using the termination condition in step 2 and the inclusion in Lemma 3.2(d), we conclude that the output of AIDAL is a (ρ, η)-stationary point of (1). $\square$

4 Numerical Experiments

This section examines the performance of the AIDAL method for solving problems of the form given in (1). It contains four subsections. The first three contain the following problem classes: (i) a class of linearly-constrained quadratic programming problems considered in [10]; (ii) the sparse principal component analysis (PCA) problem in [5]; and (iii) a class of linearly-constrained quadratic matrix problems considered in [12, 13]. The last subsection gives a few comments about the results.

Before proceeding with the results, we describe the implementation details of our algorithms and the setup of our experiments. These include specific parameter choices, special modifications, and added heuristics.

We first discuss the three implementation of the AIDAL method, labeled rADL0, rADL1, and tADL1 considered in this section. Broadly speaking, tADL1 is an implementation of the theoretical version of AIDAL in Algorithm 2.1, while rADL0 and rADL1 are implementations of an adaptive/relaxed version of AIDAL in Algorithm D.1. In particular, the adaptive version of AIDAL introduces a novel line search scheme for adaptively choosing the prox parameter λ in AIDAL (for further details, see the discussion in Appendix D). In terms of parameters, each AIDAL implementation uses $p_0 = 0$, $c_1 = \max\{1, M/\|A\|^2\}$, and $\sigma = 0.3$ for every outer iteration of the method. However, rADL0 chooses $(\chi, \theta, \lambda_0) = (1, 0, 10)$ with a heuristic choice of $\alpha_{\chi,\theta} = 0$ and $a_\theta = 1$ in the definition of Ψ_i^θ, while rADL1 and tADL1 choose $(\chi, \theta) = (1/6, 1/2)$ and $\lambda_0 = 10$ for rADL. Note that rADL0 uses parameters that do not satisfy (12), but work well in practice.

Besides the above AIDAL implementations, we also use four other methods as benchmarks. The first one, named iALM, is an implementation of the inexact proximal augmented Lagrangian method of [18] in which: (i) its key parameters are

$$\sigma = 5, \quad \beta_0 = \max\left\{1, \frac{\max\{m, M\}}{\|A\|^2}\right\}, \quad w_0 = 1, \quad \boldsymbol{y}^0 = 0, \quad \gamma_k = \frac{(\log 2)\, \|Ax^1\|}{(k+1)\, [\log(k+2)]^2},$$

for every $k \geq 1$; and (ii) the starting point given to the k^{th} APG call is set to be $\boldsymbol{x}^{k-1}$, which is the prox center for the k^{th} prox subproblem. The second one, named IPL, is an implementation of the inexact proximal augmented Lagrangian method of [13, Section 5] where: (i) c_k is doubled in its step 4 rather than quintupled; and (ii) $\sigma = 0.3$. The third one, named QP, is a practical modification of the quadratic penalty method of [10] in which: (i) each ACG subproblem in step 1 of the AIPP method is stopped when the condition

$$\|u_j\| + 2\eta_j \leq \sigma\|x_0 - x_j + u_j\|^2$$

holds; and (ii) it uses the parameters $\sigma = 0.3$ and $c = \max\{1, M/\|A\|^2\}$. The fourth and last one, named RQP, is an instance of the relaxed quadratic penalty method of [12] in which: (i) it uses the AIPPv1 variant described in [12, Section 6] with the parameters $(\theta, \tau) = (4, 10[\lambda_0 M + 1])$ and $\lambda_0 = 10$; and (ii) it uses the initial penalty parameter $c_1 = \max\{1, M/\|A\|^2\}$. It is also worth mentioning that every method except the iALM replaces its ACG prox subproblem solver by a more practical FISTA variant whose key iterates are as described in [24] and whose main stepsize parameter is adaptively estimated by a line search subroutine described in [8, Algorithm 5.2.1].

We now give some comments about the benchmark algorithms. First, iALM differs from the other tested methods in that it uses an ACG variant with a termination criterion that is different from the one in (39) and/or its relaxation. Second, the main difference between the AIDAL variants and IAIPAL methods is in how they decide when to double c_k, i.e., step 3 of Algorithm 2.1. In particular, the condition used in the IAIPAL method depends on both σ and k whereas the condition in the AIDAL variants do not. Finally, QP-AIPP is the only method that can be run without requiring any regularity conditions on the linear constraint and without assuming that $D_h < \infty$. In Table 4.1, we summarize the adaptivity of the above methods in terms of the adaptivity of the curvature constants M and m in assumption (A2). In particular, we consider the adaptivity of m to be equivalent to the adaptivity of the prox stepsize λ.

Properties	rADL0	rADL1	tADL1	iALM	IPL	QP	RQP
Estimates M	✓	✓	✓	✘	✓	✓	✓
Estimates m	✓	✓	✘	✘	✘	✘	✓

Table 4.1: The first (resp. second) row indicates whether a line search is used to estimate the curvature constant M (resp. m) in assumption (A2) for a prox subproblem. Note that estimation of m is equivalent to estimation of the prox stepsize λ.

For a linear operator A, a proper lower semicontinuous convex function h, a function f satisfying assumptions (A2)–(A4), a tolerance pair $(\rho, \eta) \in \mathbb{R}^2_{++}$, and an initial point $z_0 \in \operatorname{dom} h$, each of the methods of this section seeks a pair $([\hat{z}, \hat{p}], \hat{v})$ satisfying

$$\hat{v} \in \nabla f(\hat{z}) + \partial h(\hat{z}) + A^*\hat{p}, \quad \frac{\|\hat{v}\|}{\|\nabla f(z_0)\| + 1} \leq \rho, \quad \frac{\|A\hat{z} - b\|}{\|Az_0 - b\| + 1} \leq \eta. \tag{32}$$

In particular, the quadratic programming and matrix problem experiments consider $(\rho, \eta) = (10^{-3}, 10^{-3})$, while the sparse PCA experiments consider $(\rho, \eta) = (10^{-4}, 10^{-4})$. Moreover, defining c_0 to be the initial penalty parameter and n_i to be the number of outer iterations with $c = c_0 2^i$, we also report the following metrics:

$$c_{\text{wavg}} := \frac{\sum_{i\geq 0} n_i \cdot c_0 2^i}{\sum_{i\geq 0} n_i}, \quad c_{\max} := \text{ final penalty parameter } c.$$

All experiments are implemented in MATLAB 2020b and are run on Linux 64-bit machines, each containing Xeon E5520 processors and at least 8 GB of memory. Furthermore, the bold numbers in each of the tables of this section indicate the method that performed the most efficiently for a given benchmark, e.g., runtime or (innermost) iteration count. Finally, it is worth mentioning that the code for replicating these experiments is freely available online[2].

4.1 Linearly-Constrained Quadratic Programming

Given a pair of dimensions $(l,n) \in \mathbb{N}^2$, scalar pair $(\alpha_1,\alpha_2) \in \mathbb{R}^2_{++}$, matrices $A,B,C \in \mathbb{R}^{l\times n}$, positive diagonal matrix $D \in \mathbb{R}^{n\times n}$, and vector pair $(b,d) \in \mathbb{R}^l \times \mathbb{R}^l$, this subsection considers the following linearly-constrained quadratic programming (LCQP) problem:

$$\begin{aligned} \min_z \ & \frac{\alpha_1}{2}\|Cz-d\|^2 - \frac{\alpha_2}{2}\|DBz\|^2 \\ \text{s.t.} \ & Az=b, \quad z\in\Delta_n, \end{aligned}$$

where $\Delta_n = \{z \in \mathbb{R}^n_+ : \sum_{i=1}^n z_i = 1\}$ denotes the n-dimensional simplex.

We now describe the experiment parameters for the instances considered. First, the dimensions are set to $(l,n)=(10,50)$ and *all* of the entries in A, B, and C are nonzero. Second, the entries of A,B,C,b, and d (resp., D) are generated by sampling from the uniform distribution $\mathcal{U}[0,1]$ (resp., $\mathcal{U}[1,1000]$). Third, the initial starting point z_0 is generated by sampling a random vector $\tilde{z}_0$ from $\mathcal{U}^2[0,1]$ and setting $z_0 = \tilde{z}_0/\|\tilde{z}_0\|$. Fourth, using the well-known fact that $\|z\| \le 1$ for every $z\in\Delta_n$, the auxiliary parameters for the iALM are $B_i = \|a_i\|$, $L_i = 0$, and $\rho_i = 0$, for every i, where a_i is the i^{th} row of A. Finally, the composite form of the problem is

$$f(z) = \frac{\alpha_1}{2}\|Cz-d\|^2 - \frac{\alpha_2}{2}\|DBz\|^2, \quad h(z) = \delta_{\Delta_n}(z),$$

and each problem instance uses a scalar pair $(\alpha_1,\alpha_2) \in \mathbb{R}^2_{++}$ so that $M = \lambda_{\max}(\nabla^2 f)$ is a particular value given in the table below and $m = -M/3$.

We now present the numerical results for this set of problem instances in Table 4.2 and Table 4.3.

M	Iteration Count							Runtime (seconds)						
	rADL0	rADL1	tADL1	iALM	IPL	QP	RQP	rADL0	rADL1	tADL1	iALM	IPL	QP	RQP
10^2	**958**	1196	6910	11498	26256	20473	2455	**2.0**	2.5	14.0	13.8	53.4	37.9	4.6
10^3	2538	2807	7307	12669	25846	20354	**2261**	5.2	5.7	15.8	17.1	53.9	38.2	**4.2**
10^4	**856**	2624	7307	12729	25846	20497	2710	**1.7**	5.4	15.2	15.8	53.0	38.4	5.0
10^5	**908**	2649	7322	12743	25846	20311	4571	**1.8**	5.3	14.7	15.0	52.6	38.5	8.8
10^6	**1045**	2514	7322	12744	25846	20313	7889	**2.1**	5.2	15.2	15.8	60.0	39.9	14.8

Table 4.2: Innermost iteration counts and runtimes for LCQP problems.

It is worth mentioning that we also attempted to add the sProxALM method of [30, 31] to our list of benchmark methods with its penalty parameter set to $\Gamma = 10$ and all other parameters set as in [30, Algorithm 2]. However, for every problem instance, sProxALM failed to obtain a solution as in (32) under a generous time limit of 3600 seconds, so we have excluded its addition to the results above. Note that we did not test sProxALM on the other numerical experiments because their settings did not fall into settings considered by [30, 31] (i.e., where the composite function h needs to be the indicator function for a polyhedral set). Also, contrary to our AIDAL implementations, [30, 31] does not provide a concrete way of choosing the parameters (adaptively or otherwise) of sProxALM to ensure its convergence.

[2]See `https://github.com/wwkong/nc_opt/tree/master/tests/papers/aidal`.

M	$c_{\max}$							$c_{\text{wavg}}/c_{\max}$						
	rADL0	rADL1	tADL1	iALM	IPL	QP	RQP	rADL0	rADL1	tADL1	iALM	IPL	QP	RQP
10^2	**6E+1**	2E+3	2E+3	3E+3	3E+5	4E+3	4E+3	0.10	0.15	0.02	0.02	0.75	0.20	0.08
10^3	**2E+3**	4E+4	4E+4	3E+4	3E+6	4E+4	4E+4	0.12	0.14	0.01	0.02	0.75	0.19	0.10
10^4	**2E+4**	4E+5	4E+5	3E+5	3E+7	4E+5	4E+5	0.18	0.13	0.01	0.02	0.75	0.20	0.13
10^5	**2E+5**	4E+6	4E+6	3E+6	3E+8	4E+6	4E+6	0.18	0.13	0.01	0.02	0.75	0.19	0.14
10^6	**2E+6**	4E+7	4E+7	3E+7	3E+9	4E+7	4E+7	0.18	0.13	0.01	0.02	0.75	0.19	0.15

Table 4.3: Penalty parameter statistics for LCQP problems.

4.2 Sparse PCA

Given integer k, positive scalar pair $(\nu, b) \in \mathbb{R}^2_{++}$, and matrix $\Sigma \in S^n_+$, this subsection considers the following sparse principal component analysis (SPCA) problem:

$$\begin{aligned} \min_{\Pi,\Phi} \ & \langle \Sigma, \Pi \rangle_F + \sum_{i,j=1}^{n} q_\nu(\Phi_{ij}) + \nu \sum_{i,j=1}^{n} |\Phi_{ij}| \\ \text{s.t.} \ & \Pi - \Phi = 0, \quad (\Pi, \Phi) \in \mathcal{F}^k \times \mathbb{R}^{n\times n}, \end{aligned}$$

where $\mathcal{F}^k = \{z \in S^n_+ : 0 \preceq z \preceq I, \operatorname{tr} M = k\}$ denotes the k–Fantope and $q_\nu(\cdot) + \nu|\cdot|$ is the minimax concave penalty (MCP) function given by

$$q_\nu(t) := \begin{cases} -t^2/(2b), & \text{if } |t| \le b\nu, \\ b\nu^2/2 - \nu|t|, & \text{if } |t| > b\nu, \end{cases} \quad \forall t \in \mathbb{R}.$$

Note that the effective domain of this problem is unbounded, and hence, only the QP method is guaranteed to converge to an approximate stationary point in general.

We now describe the experiment parameters for the instances considered. First, the scalar parameters are chosen to be $(\nu, b) = (100, 0.005)$. Second, the matrix Σ is generated according to an eigenvalue decomposition $\Sigma = P\Lambda P^T$, based on a parameter pair (s, k), where k is as in the problem description and s is a positive integer. In particular, we choose $\Lambda = (100, 1, ..., 1)$, the first column of P to be a sparse vector whose first s entries are $1/\sqrt{s}$, and the other entries of P to be sampled randomly from the standard Gaussian distribution. Third, the initial starting point is $(\Pi_0, \Phi_0) = (D_k, 0)$ where D_k is a diagonal matrix whose first k entries are 1 and whose remaining entries are 0. Fourth, the curvature parameters for each problem instance are $m = M = 1/b$ and k is fixed at $k = 1$. Fifth, for the iALM, we make the following parameter choices based on a relaxed (but unverified) assumption that its generated iterates lie in $\mathcal{F}_k \times \mathcal{F}_k$: $B_i = 1$, $L_i = 0$, and $\rho_i = 0$ for all i. Sixth, the composite form of the problem is

$$f(\Pi, \Phi) = \langle \Sigma, \Pi \rangle_F + \sum_{i,j=1}^{n} q_\nu(\Phi_{ij}), \quad h(\Pi, \Phi) = \delta_{\mathcal{F}^k}(\Pi) + \nu \sum_{i,j=1}^{n} |\Phi_{ij}|,$$
$$A(\Pi, \Phi) = \Pi - \Phi, \quad b = 0,$$

and each problem instance considers a different value of s.

We now present the numerical results for this set of problem instances in Tables 4.4 and 4.5.

s	Iteration Count					Runtime (seconds)				
	rADL0	iALM	IPL	QP	RQP	rADL0	iALM	IPL	QP	RQP
5	**394**	44952	2779	22559	2990	**3.0**	139.2	17.0	118.1	16.6
10	**403**	47373	2646	19984	2983	**2.7**	143.1	14.8	103.8	15.8
15	**398**	45552	2628	20126	2996	**2.4**	138.2	15.1	103.8	16.6

Table 4.4: Innermost iteration counts and runtimes for SPCA problems.

s	$c_{\max}$					$c_{\text{wavg}}/c_{\max}$				
	rADL0	iALM	IPL	QP	RQP	rADL0	iALM	IPL	QP	RQP
5	**6E+3**	4E+6	3E+5	4E+6	2E+6	0.57	0.03	0.33	0.04	0.09
10	**6E+3**	4E+6	3E+5	4E+6	2E+6	0.57	0.03	0.28	0.03	0.09
15	**6E+3**	4E+6	3E+5	4E+6	2E+6	0.57	0.03	0.35	0.03	0.09

Table 4.5: Penalty parameter statistics for SPCA problems.

4.3 Linearly-Constrained Quadratic Matrix Problem

Given a pair of dimensions $(l, n) \in \mathbb{N}^2$, scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$, linear operators $\mathcal{A} : S^n_+ \mapsto \mathbb{R}^l$, $\mathcal{B} : S^n_+ \mapsto \mathbb{R}^n$, and $\mathcal{C} : S^n_+ \mapsto \mathbb{R}^l$ defined by

$$[\mathcal{A}(z)]_i = \langle A_i, z \rangle, \quad [\mathcal{B}(z)]_j = \langle B_j, z \rangle, \quad [\mathcal{C}(z)]_i = \langle C_i, z \rangle,$$

for matrices $\{A_i\}_{i=1}^l, \{B_j\}_{j=1}^n, \{C_i\}_{i=1}^l \subseteq \mathbb{R}^{n\times n}$, positive diagonal matrix $D \in \mathbb{R}^{n\times n}$, and vector pair $(b, d) \in \mathbb{R}^l \times \mathbb{R}^l$, this subsection considers the following linearly-constrained quadratic matrix (LCQM) problem:

$$\begin{aligned} \min_z \ & \frac{\alpha_1}{2}\|\mathcal{C}(z) - d\|^2 - \frac{\alpha_2}{2}\|D\mathcal{B}(z)\|^2 \\ \text{s.t. } & \mathcal{A}(z) = b, \quad z \in P_n, \end{aligned}$$

where $P_n = \{z \in S^n_+ : \operatorname{tr} z = 1\}$ denotes the n-dimensional spectraplex.

We now describe the experiment parameters for the instances considered. First, the dimensions are set to $(l, n) = (20, 100)$ and only 1.0% of the entries of the submatrices A_i, B_j, and C_i are nonzero. Second, the entries of A_i, B_j, C_i, b, and d (resp., D) are generated by sampling from the uniform distribution $\mathcal{U}[0, 1]$ (resp., $\mathcal{U}[1, 1000]$). Third, the initial starting point z_0 is a random point in S^n_+. More specifically, three unit vectors $\nu_1, \nu_2, \nu_3 \in \mathbb{R}^n$ and three scalars $e_1, e_2, e_2 \in \mathbb{R}_+$ are first generated by sampling vectors $\tilde{\nu}_i \sim \mathcal{U}^n[0, 1]$ and scalars $\tilde{d}_i \sim \mathcal{U}[0, 1]$ and setting $\nu_i = \tilde{\nu}_i/\|\tilde{\nu}_i\|$ and $e_i = \tilde{e}_i/(\sum_{j=1}^3 \tilde{e}_i)$ for $i = 1, 2, 3$. The initial iterate for the first subproblem is then set to $z_0 = \sum_{i=1}^3 e_i\nu_i\nu_i^T$. Fourth, using the well-known fact that $\|z\|_F \leq 1$ for every $z \in P_n$, the auxiliary parameters for the iALM are

$$B_i = \|A_i\|_F, \quad L_i = 0, \quad \rho_i = 0 \quad \forall i \geq 1.$$

Finally, the composite form of the problem is

$$f(z) = \frac{\alpha_1}{2}\|\mathcal{C}(z) - d\|^2 - \frac{\alpha_2}{2}\|D\mathcal{B}(z)\|^2, \quad h(z) = \delta_{P_n}(z), \quad A(z) = \mathcal{A}(z),$$

and each problem instance uses a scalar pair $(\alpha_1, \alpha_2) \in \mathbb{R}^2_{++}$ so that $M = \lambda_{\max}(\nabla^2 f)$ is a particular value given in the table below and $m = -M/4$.

We now present the numerical results for this set of problem instances in Tables 4.6 and 4.7.

M	Iteration Count					Runtime (seconds)				
	rADL0	iALM	IPL	QP	RQP	rADL0	iALM	IPL	QP	RQP
100	**388**	66000	6863	37470	8293	**4.4**	323.3	68.7	344.6	85.6
200	**486**	70551	6902	37696	1475	**5.6**	334.9	66.9	335.4	13.4
400	**674**	72760	6902	37972	1562	**7.6**	347.5	67.9	339.0	14.2
1600	**1090**	74200	6921	38203	1309	12.6	361.4	68.9	346.3	**12.1**
3200	1400	74568	6921	38243	**1327**	16.0	369.8	74.1	352.3	**12.1**

Table 4.6: Innermost iteration counts and runtimes for LCQM problems.

M	$c_{\max}$					$c_{\text{wavg}}/c_{\max}$				
	rADL0	iALM	IPL	QP	RQP	rADL0	iALM	IPL	QP	RQP
100	**4E+1**	2E+3	6E+2	1E+3	1E+3	0.27	0.08	0.96	0.30	0.01
200	**8E+1**	3E+3	1E+3	3E+3	3E+3	0.29	0.08	0.97	0.30	0.08
400	**2E+2**	6E+3	3E+3	5E+3	5E+3	0.33	0.08	0.97	0.31	0.11
1600	**6E+2**	2E+4	1E+4	2E+4	2E+4	0.39	0.08	0.97	0.31	0.12
3200	**1E+3**	5E+4	2E+4	4E+4	4E+4	0.39	0.08	0.97	0.31	0.13

Table 4.7: Penalty parameter statistics for LCQM problems.

4.4 Comments about Numerical Experiments

Algorithm rADL0 is generally the most efficient in terms of total inner (or ACG) iterations, runtime, and final penalty parameter used. Moreover, the experiments in Subsection 4.1 demonstrate that the adaptivity of m (or equivalently λ) substantially improves AIDAL in terms of both inner iteration count and runtime. Finally, while the penalty ratio $c_{\text{wavg}}/c_{\max}$ is generally the lowest for iALM, the performance for iALM in terms of the number of inner iterations and runtime is generally the worst among the tested methods.

5 Concluding Remarks

Similar to the analyses in [18,20], the analysis of the AIDAL method strongly makes use of assumption (A3) and the assumption that $D_h < \infty$ to obtain its competitive $\mathcal{O}(\varepsilon^{-5/2}\log\varepsilon^{-1})$ iteration complexity when $\varepsilon = \rho = \eta$. However, we conjecture that the these two assumptions may be removed using the more complicated analysis in [23] to obtain a slightly worse $\mathcal{O}(\varepsilon^{-3}\log\varepsilon^{-1})$ iteration complexity (like in [23]).

Like the adaptive prox-stepsize AIDAL in Appendix D, another possible extension of AIDAL is one in which λ, χ, and θ are simultaneously chosen in an adaptive manner. Moreover, it would be interesting to develop an adaptive AIDAL (as above) that has the same iteration complexity bound as the nonadaptive AIDAL in Algorithm 2.1.

A Key Technical Bounds

The appendix presents several key technical bounds that are used in the analysis of AIDAL.

The result below presents a technical inequality that is useful in establishing the bound on $\|\hat{v}_i\|$ in Proposition 3.1(b).

Lemma A.1. *For every $(\tau, \theta) \in [0,1]^2$ satisfying $\tau \le \theta^2$ and every $a, b \in \mathbb{R}^n$, we have that*

$$\|a - (1-\theta)b\|^2 - \tau\|a\|^2 \ge \left[\frac{(1-\tau) - (1-\theta)^2}{2}\right]\left(\|a\|^2 - \|b\|^2\right). \tag{33}$$

Proof. Let $a, b \in \mathbb{R}^n$ be fixed and define

$$z = \begin{bmatrix} \|a\| \\ \|b\| \end{bmatrix}, \quad M = \begin{bmatrix} (1-\tau) + (1-\theta)^2 & -2(1-\theta) \\ -2(1-\theta) & (1-\tau) + (1-\theta)^2 \end{bmatrix}. \tag{34}$$

Moreover, using our assumption of $\tau \le \theta^2 \le 1$, observe that

$$\begin{aligned} \det M &= \left[(1-\tau) + (1-\theta)^2 - 2(1-\theta)\right]\left[(1-\tau) + (1-\theta)^2 + 2(1-\theta)\right] \\ &= \left[\theta^2 - \tau\right]\left[(1-\tau) + (1-\theta)^2 + 2(1-\theta)\right] \ge 0, \end{aligned}$$

and hence, by Sylvester's criterion, it follows that $M \succeq 0$. Combining this fact with the Cauchy-Schwarz inequality and (34), we thus have that

$$\begin{aligned} &\|a - (1-\theta)b\|^2 - \tau\|a\|^2 \ge (1-\tau)\|a\|^2 - 2(1-\theta)\|a\| \cdot \|b\| + (1-\theta)^2\|b\|^2 \\ &= \frac{1}{2}z^T M z + \left[\frac{(1-\tau) - (1-\theta)^2}{2}\right]\left(\|a\|^2 - \|b\|^2\right) \ge \left[\frac{(1-\tau) - (1-\theta)^2}{2}\right]\left(\|a\|^2 - \|b\|^2\right). \end{aligned}$$ □

We now prove Proposition 2.1, which describes a bound on the Lagrange multipliers. Its techniques are a modification of the ones in [14].

We first give a few preliminary results. The first result, whose proof can be found in [4, Lemma 1.3], presents a relationship between elements in the image of a linear operator.

Lemma A.2. *For every $S \in \mathbb{R}^{m\times n}$ and $u \in \operatorname{Im} S$, we have $\sigma_S^+\|u\| \le \|Su\|$.*

The proof of the next two results can be found in [14].

Lemma A.3. *Suppose $\psi \in \overline{\operatorname{Conv}}\ \mathbb{R}^n$ is K_ψ-Lipschitz. Then, for every $y \in \operatorname{dom}\psi$, we have*

$$\partial\psi(z) \subseteq \mathcal{B}(0; K_\psi) + N_{\operatorname{dom}\psi}(z).$$

Lemma A.4. *Suppose $\psi \in \overline{\operatorname{Conv}}\ \mathbb{R}^n$ is K_ψ-Lipschitz continuous with finite diameter D_ψ. Then, for every $y, \bar{y} \in \operatorname{dom} h$ and $\xi \in \partial\psi(y)$, we have*

$$\|\xi\|\operatorname{dist}_{\partial(\operatorname{dom}\psi)}(\bar{y}) \le \left[\operatorname{dist}_{\partial(\operatorname{dom}\psi)}(\bar{y}) + \|y - \bar{y}\|\right]K_\psi + \langle \xi, y - \bar{y}\rangle.$$

The next two results mirror similar ones in [14, Section 3].

Lemma A.5. *Let G_f and D_h be as in (35) and assumption (A1). Moreover, define the scalar*

$$\xi_k := \hat{v}_k - \nabla f(z_k) - A^*\hat{p}_k \quad \forall k \ge 1. \tag{35}$$

Then, the following statements hold for every $k \ge 1$:

(a) $\xi_k \in \partial h(z_k)$;

(b) it holds that

$$\|\hat{p}_k\| \le \frac{1}{\sigma_A^+}\left[\|\xi_k\| + G_f + \frac{(1+\sigma)D_h}{\lambda}\right].$$

Proof. (a) This follows immediately from Lemma 3.2(d) and the definition of ξ_i.

(b) Using the definitions of ξ_i and G_f, the triangle inequality, part (b), and Lemma A.2 with $S = A^*$ and $u = \hat{p}_k$ yields

$$\begin{aligned}\|\hat{p}_k\| &\le \frac{\|A^*\hat{p}_k\|}{\sigma_A^+} = \frac{\|\hat{v}_k - \nabla f(z_k) - \xi_k\|}{\sigma_A^+} \le \frac{\|\xi_k\| + \|\nabla f(z_k)\| + \|\hat{v}_k\|}{\sigma_A^+} \\ &\le \frac{1}{\sigma_A^+}\left[\|\xi_k\| + \|\nabla f(z_k)\| + \frac{(1+\sigma)\|\Delta z_k\|}{\lambda}\right] \le \frac{1}{\sigma_A^+}\left[\|\xi_k\| + G_f + \frac{(1+\sigma)D_h}{\lambda}\right]. \qquad \square\end{aligned}$$

Lemma A.6. *Let $(\beta_\lambda, \bar{d})$ be as in (13). Then, the following statements hold for every $(\chi, \theta) \in (0,1)^2$ and $k \ge 1$:*

(a) $\|p_k\| \le \chi\|\hat{p}_k\| + (1-\chi)(1-\theta)\|p_{k-1}\|$;

(b) $c_k^{-1}\|\hat{p}_k\|^2 + \bar{d}\sigma_A^+\|\hat{p}_k\| \le c_k^{-1}(1-\theta)\langle \hat{p}_k, p_{k-1}\rangle + \beta_\lambda$.

Proof. (a) Using the definitions of p_k and $\hat{p}_k$ with the triangle inequality yields

$$\|p_k\| = \|\chi\hat{p}_k + (1-\chi)(1-\theta)p_{k-1}\| \le \chi\|\hat{p}_k\| + (1-\chi)(1-\theta)\|p_{k-1}\|.$$

(b) Let ξ_k, $(G_f, \bar{d})$, and D_h be as in (35), (13), and assumption (A1), respectively. Using Lemma A.5(a), the definition of $\bar{d}$, and Lemma A.4 with $(\psi, K_\psi, D_\psi) = (h, K_h, D_h)$ and $(y, \bar{y}, \varepsilon) = (z_k, \bar{z}, \delta_k)$, we have that

$$\bar{d}\|\xi_k\| \le (\bar{d} + D_h)K_h + \langle \xi_k, z_k - \bar{z}\rangle. \tag{36}$$

Moreover, the definitions of $\hat{p}_k$ and ξ_k, the fact that $z_k, \bar{z} \in \mathcal{H}$ and $A\bar{z} = b$, and the Cauchy-Schwarz inequality imply that

$$\begin{aligned}\langle \xi_k, z_k - \bar{z}\rangle &= \langle \hat{v}_k - \nabla f(z_k) - A^*\hat{p}_k, z_k - \bar{z}\rangle \\ &\le (\|\hat{v}_k\| + \|\nabla f(z_k)\|)\|z_k - \bar{z}\| - \langle \hat{p}_k, Az_k - b\rangle \\ &\le \left[\frac{(1+\sigma)D_h}{\lambda} + G_f\right]D_h + \left(\frac{1-\theta}{c_k}\right)\langle \hat{p}_k, p_{k-1}\rangle - \frac{1}{c_k}\|\hat{p}_k\|^2.\end{aligned} \tag{37}$$

Using Lemma A.5(b), (36), (37), and the definition of β_λ in (13), we thus conclude that

$$\begin{aligned}\frac{1}{c_k}\|\hat{p}_k\|^2 + \bar{d}\sigma_A^+\|\hat{p}_k\| &\le \frac{1}{c_k}\|\hat{p}_k\|^2 + \bar{d}\|\xi_k\| + \left[G_f + \frac{(1+\sigma)D_h}{\lambda}\right]\bar{d} \\ &\le \langle \xi_k, z_k - \bar{z}\rangle + \frac{\beta_\lambda \bar{d}}{\bar{d} + D_h} \le \left(\frac{1-\theta}{c_k}\right)\langle \hat{p}_k, p_{k-1}\rangle + \beta_\lambda. \qquad \square\end{aligned}$$

We are now ready to give the proof of Proposition 2.1.

Proof of Proposition 2.1. We proceed by induction on k. Since $B_p \ge \|p_0\|$, the desired bound trivially holds for $k = 0$. Assume now that $\|p_k\| \le B_p$ holds for some $k \ge 0$. If $\|\hat{p}_{k+1}\| = 0$, then clearly

$$\|p_{k+1}\| \le \chi\|\hat{p}_{k+1}\| + (1-\chi)(1-\theta)\|p_k\| = (1-\chi)(1-\theta)B_p \le B_p,$$

so suppose that $\|\hat{p}_{k+1}\| > 0$. Using Lemma A.6(b), the Cauchy-Schwarz inequality, and the induction hypothesis we have that

$$\begin{aligned}\left[\bar{d} + \frac{1}{c_{k+1}\sigma_A^+}\|\hat{p}_{k+1}\|\right]\|\hat{p}_{k+1}\| &= \frac{1}{\sigma_A^+}\left[\frac{1}{c_{k+1}}\|\hat{p}_{k+1}\|^2 + \bar{d}\sigma_A^+\|\hat{p}_{k+1}\|\right] \le \frac{1}{\sigma_A^+}\left[\left(\frac{1-\theta}{c_{k+1}}\right)\langle \hat{p}_{k+1}, p_k\rangle + \beta_\lambda\right] \\ &= \frac{1}{c_{k+1}\sigma_A^+}\left[c_{k+1}\beta_\lambda + (1-\theta)\|p_k\| \cdot \|\hat{p}_{k+1}\|\right] \le \left[\bar{d} + \frac{1}{c_{k+1}\sigma_A^+}\|\hat{p}_{k+1}\|\right]B_p.\end{aligned}$$

and hence that $\|\hat{p}_{k+1}\| \le B_p$. Combining this bound with the induction hypothesis and part (a), we finally conclude that

$$\|p_{k+1}\| \le \chi\|\hat{p}_{k+1}\| + (1-\chi)(1-\theta)\|p_k\| \le B_p. \qquad \square$$

B Statement and Properties of the ACG Algorithm

Recall from Section 1 that our interest is in solving (1) by inexactly solving NCO subproblems of the form in (3). This subsection presents an ACG algorithm for inexactly solving latter type of problem and it considers the more general class of NCO problems

$$\min_{u\in\mathbb{R}^n} \left\{\psi_s(u) + \psi_n(u)\right\}, \tag{38}$$

where the functions ψ_s and ψ_n are assumed to satisfy the following assumptions:

(B1) $\psi_n : \mathbb{R}^n \mapsto (-\infty, \infty]$ is a proper closed convex function.

(B2) ψ_s is convex and continuously differentiable on $\mathbb{R}^n$ and satisfies

$$\frac{\mu}{2}\|z' - z\|^2 \le \psi_s(z') - \ell_{\psi_s}(z'; z) \le \frac{L}{2}\|z' - z\|^2$$

for every $z', z \in \mathbb{R}^n$ and some $L > 0$ and $\mu \in (0, L]$.

Clearly, problem (3) is a special case of (38), and hence, any result that is stated in the context of (38) also applies to (3).

The pseudocode for the ACG algorithm is stated in Algorithm B.1 which, for a given a pair $(\sigma, x_0) \in \mathbb{R}_{++} \times \operatorname{dom}\psi_n$, inexactly solves (38) by obtaining a pair (z, v) satisfying

$$v \in \nabla\psi_s(z) + \partial\psi_n(z), \quad \|v\| \le \sigma\|z - x_0\|. \tag{39}$$

Note that if ACG algorithm obtains the aforementioned triple with $\sigma = 0$ then the first component of the triple is, in fact, a global solution of (38). Indeed, if $\sigma = 0$ then the above inequality implies that $v = 0$, and the above inclusion reduces to $0 \in \partial(\psi_s + \psi_n)(z)$, which in view of (7) clearly implies that z is a global solution of (38).

The result below presents some basic properties about the ACG algorithm.

Proposition B.1. *The following properties hold about the ACG algorithm:*

(a) for every $j \ge 0$, it holds that $u_{j+1} \in \nabla\psi_s(x_{j+1}) + \psi_n(x_{j+1})$;

(b) for every $\sigma > 0$, it stops and outputs a pair (z, v) satisfying (39) in a number of iterations bounded above by

$$\left\lceil 1 + 2\sqrt{\frac{2L}{\mu}} \log_1^+ \left\{\frac{4L^2}{\mu}\left(\frac{1}{\mu} + \frac{36L}{\sigma^2}\right)\right\}\right\rceil.$$

Proof. (a) See [9, Lemma 3.3(c)] with $(\psi^s, \psi^n, y_{j+1}) = (\psi_s, \psi_n, x_{j+1})$.

(b) See [9, Proposition 3.5] with $(\overline{L}, \theta) = (L, +\infty)$ and $(\psi^s, \psi^n, y_{j+1}) = (\psi_s, \psi_n, x_{j+1})$. $\square$

Algorithm B.1: Accelerated Composite Gradient (ACG) Algorithm

Input : $(\sigma, x_0) \in (0,1) \times \operatorname{dom} \psi_n$.
Output: a pair $(z, v) \in \operatorname{dom} \psi_n \times \mathbb{R}^n$ satisfying (39).

Function `ACG`$(\{\psi_s, \psi_n\}, \{L, \mu\}, \sigma, x_0)$**:**
  `STEP 0` (initialization):
  Set $y_0 \leftarrow x_0$, $A_0 \leftarrow 0$.
  for $j \leftarrow 0, 1, ...$ **do**
    `STEP 1` (main iterates):
    find the positive scalar a_j satisfying $a_j^2 = \frac{(1+\mu A_j)(a_j + A_j)}{L}$
    $A_{j+1} \leftarrow A_j + a_j$
    $\tilde{x}_j \leftarrow \frac{A_j}{A_{j+1}} x_j + \frac{A_{j+1} - A_j}{A_{j+1}} y_j$
    $x_{j+1} \leftarrow \operatorname{argmin}_{y \in \mathbb{R}^n} \left\{ \ell_{\psi_s}(y; \tilde{x}_j) + \psi_n(y) + \frac{L+\mu}{2} \|y - \tilde{x}_j\|^2 \right\}$
    $y_{j+1} \leftarrow y_j + \frac{a_j}{1+\mu A_{j+1}} [L(x_{j+1} - \tilde{x}_j) + \mu(x_{j+1} - y_j)]$
    `STEP 2` (termination check):
    $u_{j+1} \leftarrow \nabla\psi_s(x_{j+1}) - \nabla\psi_s(\tilde{x}_j) + (L+\mu)(\tilde{x}_j - x_{j+1})$
    if $\|u_{j+1}\| \le \sigma \|x_{j+1} - x_0\|$ **then**
      return (x_{j+1}, u_{j+1})

C Necessary Optimality Conditions

This appendix shows that if $\hat{z}$ local minimum of (1) then condition (11) holds. Throughout this appendix, we denote

$$\psi'(x; d) = \lim_{t \downarrow 0} \frac{\psi(x + td) - \psi(x)}{t}$$

as the directional derivative of a function ψ at x in the direction d.

The first useful result presents a relationship between directional derivatives of composite functions and the usual first-order necessary conditions.

Lemma C.1. *Let $g : \mathbb{R}^n \mapsto (-\infty, \infty]$ be a proper convex function, and let f be a differentiable function on $\operatorname{dom} g$. Then, for every $x \in \operatorname{dom} g$, the following statements hold:*

(a) $\inf_{\|d\| \le 1} (f+g)'(x; d) = -\inf_{u \in \mathbb{R}^n} \{\|u\| : u \in \nabla f(x) + \partial g(x)\}$;

(b) if x is a local minimum of $f + h$ then $x \in \nabla f(x) + \partial h(x)$.

Proof. (a) See [15, Lemma 15] with $(\mathcal{X}, h) = (\mathbb{R}^n, g)$.

(b) This follows immediately from (a) and the fact that $(f+h)'(x; d) \ge 0$ for every $d \in \mathbb{R}^n$. □

We now establish the aforementioned necessary condition.

Proposition C.2. *Let (f, h, A, b) be as in (A1)-(A4). If $\hat{z}$ is a local minimum of* (1)*, then there exists a multiplier $\hat{p}$ such that* (11) *holds.*

Proof. We first establish an important technical identity. Let $S = \{z \in \mathbb{R}^n : Az = b\}$, let δ_S denote the indicator function of S, i.e., the function that takes value 0 if its input is in S and $+\infty$

otherwise, and let $\operatorname{ri} X$ denote the relative interior of a set X. Since assumptions (A3)–(A4) imply that $\operatorname{ri} \mathcal{H} \cap \operatorname{ri} S = \operatorname{int} \mathcal{H} \cap S \neq \emptyset$, it follows from [27, Theorem 23.8] that for every $x \in \mathcal{H} \cap S$ we have

$$\partial(\delta_S + h)(x) = \partial \delta_S(x) + \partial h(x) = N_S(x) + \partial h(x) = \{\xi + A^* p : \xi \in \partial h(x)\}. \tag{40}$$

The conclusion follows from the above identity and Lemma C.1(b) with $g = h + \delta_S$. □

D Adaptive AIDAL

This appendix presents an adaptive version of AIDAL where we choose the prox stepsize adaptively.

Before presenting the algorithm, we first motivate its construction under the assumption that the reader is familiar with the notation and results of Section 3. To begin, the careful reader may notice that the special choice of $\lambda = 1/(2m)$ in AIDAL (Algorithm 2.1) is only needed to ensure that the function $\lambda \mathcal{L}_c^\theta(\cdot; p) + \|\cdot\|^2$ is strongly convex with respect to the norm $\|x\|_c = \langle x, [(1-\lambda m)I + c\lambda A^* A]x \rangle$ for every $c > 0$ and $p \in A(\mathbb{R}^n)$. Moreover, this global property is only needed to show that:

(i) the k^{th} ACG call of AIDAL stops with a pair (z_k, v_k) satisfying $\|v_k\| \le \sigma \|z_k - z_{k-1}\|$;

(ii) $\lambda \|\hat{v}_i\| \lesssim \Psi_{k-1}^\theta - \Psi_k^\theta$.

The other technical details of Section 3, such as the boundedness of Ψ_i^θ, are straightforward to show as long as the prox stepsize is bounded. As a consequence, a natural relaxation of AIDAL is to employ a line search at its k^{th} outer iteration for the largest λ within a bounded range satisfying conditions (i) and (ii) above.

In Algorithm D.1, we present one possible relaxation. Specifically, the k^{th} prox stepsize λ_k is chosen from a set of candidates in the range $(0, \lambda_{k-1}]$.

Algorithm D.1: Adaptive AIDAL Method

Input : Same as in Algorithm 2.1 but with additional parameters $\gamma > 1$ and $\lambda_0 > 0$.
Output: Same as in Algorithm 2.1.

Function `AdapAIDAL`(M,$\{\sigma, \chi, \theta, \lambda_0\}, \{c_1, z_0, p_0\}, \{\rho, \eta\}, \gamma$)**:**
  $\lambda_0 \leftarrow \lambda$
  for $k \leftarrow 1, 2, ...$ **do**
    find the smallest nonnegative integer β_k such that the ACG call in step 1 of Algorithm 2.1 with $\lambda = \gamma^{-\beta_k} \lambda_{k-1}$ stops with a pair (z_k, v_k) satisfying

$$\begin{cases} \|v_k\| \le \sigma \|z_k - z_{k-1}\| & \text{if } k \ge 1, \text{ and} \\ \|v_k + z_{k-1} - z_k\|^2 \le 32\lambda(\Psi_{k-1}^\theta - \Psi_k^\theta) & \text{if } k \ge 2, \end{cases} \tag{41}$$

    where Ψ_k^θ is given in (23)
    set $\lambda_k \leftarrow \gamma^{-\beta_k} \lambda_{k-1}$
    execute steps 1–4 of Algorithm 2.1 with $\lambda = \lambda_k$

We now make a few remarks about Algorithm D.1. First, the candidate search space for the k^{th} prox stepsize forms a geometrically decreasing sequence and $\lambda_k \le \lambda_{k-1}$. Second, the first condition of (41) corresponds to condition (i), while the second condition corresponds to condition (ii). Moreover, the second condition of (41) always holds when $\lambda = 1/(2m)$ due to Lemma 3.3, Lemma 3.4, and the definition of $\hat{v}_i$ which imply (cf. the proof of Proposition 3.1(b)) that

$$\|v_k + z_{k-1} - z_k\|^2 = \lambda^2\|\hat{v}_k\|^2 \leq 32\lambda(\Psi^\theta_{k-1} - \Psi^\theta_k).$$

Third, in view of the previous remark, since conditions (i) and (ii) are always satisfied whenever $\lambda \leq 1/(2m)$, we also have that $\lambda_k \in [1/(2\gamma m), \lambda_0]$ and, hence, the sequence $\{\lambda_k\}_{k\geq 1}$ is bounded.

Notice that it is not immediately clear how one obtains β_k at the k^{th} outer iteration. One possible approach is to apply an adaptive ACG variant to the stepsize sequence $\{\lambda_{k-1}\beta^{-j}\}_{j\geq 0}$ in which the variant has a mechanism to determine if at least one of the conditions in (41) is reachable. This is so that if none of the conditions in (41) are reachable for some candidate λ, then the variant can be called again with a smaller stepsize. One example is the adaptive ACG variant in [9], which contains a mechanism for determining the reachability of the first condition in (41) and can even adaptively choose its other curvature parameters, such as L in Algorithm B.1. Note that if the ACG has already been called with the β_k satisfying (41) during the β_k line search, then it does not need to be called again when executing the steps of Algorithm 2.1.

Before closing this section, we briefly discuss the convergence and iteration complexity of the method. Convergence of the method is straightforward to establish using the same techniques of Section 3 and the fact that λ_k is bounded (see the remarks above). On the other hand, it can be shown that the iteration complexity of the method is on the same order of complexity as in Theorem 2.3. Without going through the cumbersome technical details, we assert that this follows from the boundedness of the stepsizes λ_k, the fact that the search for the next stepsize is done geometrically, and arguments similar to other adaptive augmented Lagrangian/penalty methods such as the one in [11].

Data Availability Statement

The data and code generated, used, and/or analyzed during the current study are publicly available in the `NC-OPT` GitHub repository[3] under the directory `./tests/papers/aidal/`.

Ethics Statement

The authors declare that they have no conflict of interest.

References

[1] N. S. Aybat and G. Iyengar. A first-order smoothed penalty method for compressed sensing. *SIAM J. Optim.*, 21(1):287–313, 2011.

[2] N. S. Aybat and G. Iyengar. A first-order augmented Lagrangian method for compressed sensing. *SIAM J. Optim.*, 22(2):429–459, 2012.

[3] D. Boob, Q. Deng, and G. Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Math. Program.*, pages 1–65, 2022.

[4] M. L. N. Goncalves, J. G. Melo, and R. D. C. Monteiro. Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. *Pac. J. Optim.*, 15(3):379–398, 2019.

[5] Q. Gu, Z. Wang, and H. Liu. Sparse pca with oracle property. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Adv. Neural Inf. Process. Syst. 27*, pages 1529–1537. Curran Associates, Inc., 2014.

[3] See https://github.com/wwkong/nc_opt.

[6] D. Hajinezhad and M. Hong. Perturbed proximal primal-dual algorithm for nonconvex nonsmooth optimization. *Math. Program.*, 176:207–245, 2019.

[7] B. Jiang, T. Lin, S. Ma, and S. Zhang. Structured nonconvex and nonsmooth optimization algorithms and iteration complexity analysis. *Comput. Optim. Appl.*, 72(3):115–157, 2019.

[8] W. Kong. Accelerated Inexact First-Order Methods for Solving Nonconvex Composite Optimization Problems. *Available on arXiv:2104.09685*, April 2021.

[9] W. Kong. Complexity-optimal and curvature-free first-order methods for finding stationary points of composite optimization problems. *arXiv preprint arXiv:2205.13055*, 2022.

[10] W. Kong, J. G. Melo, and R. D. C. Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. *SIAM J. Optim.*, 29(4):2566–2593, 2019.

[11] W. Kong, J. G. Melo, and R. D. C. Monteiro. An efficient adaptive accelerated inexact proximal point method for solving linearly constrained nonconvex composite problems. *Comput. Optim. Appl.*, 76(2):305–346, 2019.

[12] W. Kong, J. G. Melo, and R. D. C. Monteiro. An efficient adaptive accelerated inexact proximal point method for solving linearly constrained nonconvex composite problems. *Comput. Optim. Appl.*, 76(2):305–346, 2020.

[13] W. Kong, J. G. Melo, and R. D. C. Monteiro. Iteration-complexity of a proximal augmented Lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints. *Available on arXiv:2008.07080*, 2020.

[14] W. Kong, J. G. Melo, and R. D. C Monteiro. Iteration-complexity of an inner accelerated inexact proximal augmented Lagrangian method based on the classical Lagrangian function. *Available on arXiv:2008.00562*, 2020.

[15] W. Kong and R. D. C. Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM J. Optim.*, 31(4):2558–2585, 2021.

[16] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Math. Program.*, 138(1):115–139, Apr 2013.

[17] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented Lagrangian methods for convex programming. *Math. Program.*, 155(1):511–547, Jan 2016.

[18] Z. Li, P.-Y. Chen, S. Liu, S. Lu, and Y. Xu. Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization. *Int. Conf. Artif. Intell. Stat.*, pages 2170–2178, 2021.

[19] Z. Li and Y. Xu. Augmented Lagrangian–based first-order methods for convex-constrained programs with weakly convex objective. *INFORMS Journal on Optimization*, 3(4):373–397, 2021.

[20] Q. Lin, R. Ma, and Y. Xu. Inexact proximal-point penalty methods for constrained non-convex optimization. *Available on arXiv:1908.11518*, 2019.

[21] Y.-F. Liu, X. Liu, and S. Ma. On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. *Mathematics of Operations Research*, 44(2):632–650, 2019.

[22] Z. Lu and Z. Zhou. Iteration-complexity of first-order augmented Lagrangian methods for convex conic programming. *Available on arXiv:1803.09941*, 2018.

[23] J. G. Melo, R. D. C. Monteiro, and H. Wang. Iteration-complexity of an inexact proximal accelerated augmented Lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems. *Available on arXiv:2006.08048*, 2020.

[24] R. D. C. Monteiro, C. Ortiz, and B. F. Svaiter. An adaptive accelerated first-order method for convex optimization. *Comput. Optim. Appl.*, 64:31–73, 2016.

[25] I. Necoara, A. Patrascu, and F. Glineur. Complexity of first-order inexact Lagrangian and penalty methods for conic convex programming. *Optim. Methods Softw.*, pages 1–31, 2017.

[26] A. Patrascu, I. Necoara, and Q. Tran-Dinh. Adaptive inexact fast augmented Lagrangian methods for constrained convex optimization. *Optim. Lett.*, 11(3):609–626, 2017.

[27] R.T. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[28] M. Sahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[29] Y. Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Math. Program.*, 2019.

[30] J. Zhang and Z.-Q. Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *Available on arXiv:2006.16440*, 2020.

[31] J. Zhang and Z.-Q. Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM J. Optim.*, 30(3):2272–2302, 2020.